Inf. Pauta | Fortaleza, CE | v. 1 | n. 1 | jan./jun. 2016

ARTIGO

e-SCIENCE SEMÂNTICA: integração dos dados na comunicação científica

e-SCIENCE SEMANTICS: data integration in scientific communication

Elizabete Cristina de Souza de Aguiar Monteiro
UNESP
Ricardo Cesar Gonçalves Sant'Ana
UNESP
José Eduardo Santarém Segundo
UNESP/USP

RESUMO

As ciências experimentais, teóricas e computacionais estão sendo afetadas pela grande quantidade de dados coletados ou simulados, e emerge um novo contexto baseado no uso intensivo de dados, denominado por pesquisadores de *e-Science*, ou quarto paradigma da ciência. Dessa forma, tornar esse quarto paradigma em semântico, em que os dados e a literatura científica estejam integrados, linchados, disponíveis e interoperáveis, é um desafio. Novas ferramentas e linguagens são necessárias, e projetos para tornar isso possível estão sendo apresentados. Sendo assim, o artigo teve como objetivo a caracterização de três sítios: *Molecular BioSystems,* Solar-Terrestre Virtual e revista *Nature,* que têm melhorias semânticas e uso de ontologias em suas páginas, verificando a correlação entre Web semântica e *e-Science.* A metodologia utilizada consistiu no levantamento bibliográfico e revisão de literatura para discussão do tema; uso do método descritivo para descrever as características semânticas dos portais, além do método comparativo. Verificamos que a aplicação da Web semântica em portais com características da *e-Science* melhora consideravelmente a recuperação e, consequentemente, a reutilização de dados primários de pesquisa.

Palavras-chave: *e-Science*. *e-Science* semântica. Web semântica. Integração de dados.

ABSTRACT

The experimental, theoretical and computing sciences have been affected by the great quantity of data collected or simulated, and a new context based on the intensive use of data has been emerged and called by researchers as *e-Science*, or the fourth paradigm of science. Therefore, transforming this fourth paradigm in semantic, in which data and scientific literature are linked, available and interoperable, is a challenge. New tools and languages are necessary; and in order to make it possible, projects have been presented. In this regard, the paper aimed at characterizing 3 parts, *Molecular BioSystems,* Virtual Solar-Terrestrial and *Nature* magazine that presents semantic improvements and use of ontologies in its pages, checking the correlation between the Semantic Web and *e-Science.* The methodology used consisted of bibliographic survey and review of literature for discussion of the theme, descriptive method to describe the semantic features of portals, besides the comparative method. We verified that the application of semantic web in portals with *e-science* features considerably improves the recovery and consequently the reuse of research primary data.

Keywords: e-Science. Semantic e-Science. Semantic Web. Data integration.

1 INTRODUÇÃO

O desenvolvimento das Tecnologias de Informação e Comunicação (TIC) no processo de publicação e comunicação científica e como fator instrumental nos métodos da ciência proporciona transformações na práxis científica, direcionando mais atenção para o compartilhamento e reúso de dados.

A quantidade de dados gerados ou capturados por diversos aparatos tecnológicos nas várias áreas do conhecimento cresce rapidamente. A comunidade científica presencia o desenvolvimento de atividades no processo da comunicação científica, que enfatizam os aspectos relacionados à coleta, armazenamento e recuperação de grandes quantidades de dados, com destaque especial para dados em âmbito global, promovendo interesse crescente sobre *e-Science,* ou quarto paradigma da ciência.

e-Science é o ponto onde "[...] TI [tecnologia da informação] encontra cientistas" (TOLLE; TANSLEY; HEY, 2011, p. 17), onde computadores são usados para resolver problemas científicos com uso intensivo em dados, culminando em uma evolução das fases históricas da ciência denominadas pela experimentação, teoria e simulação (TOLLE; TANSLEY; HEY, 2011).

A *e-Science* é um conceito designado "[...] para as tecnologias de informação em rede de apoio às atividades de investigação científica, como a colaboração de compartilhamento de dados e divulgação dos resultados." (RIBES; LEE, 2010, p. 232). De acordo com os autores, há três aspectos que caracterizam as transformações proporcionadas pela *e-Science:* comunidade ampla e interdisciplinar de colaboração; computação aliada na coleta, representação e análise de dados; integração final (RIBES; LEE, 2010).

Na discussão sobre *e-Science,* Fox e Hendler (2011), do Instituto Politécnico Rensselaer, ressaltam três questões:

- Como esses dados, que não foram gerados por cientistas e pesquisadores, serão usados por eles?
- Como usar esses dados, que eles não produziram e nunca viram, junto aos dados que geram e usam todos os dias?
- O que se deve fazer se o cientista ou pesquisador, estudantes e não especialistas precisam dos dados de outra área do conhecimento e não conhecem os termos ou vocabulário da área?

Dados *in loco* das diversas áreas do conhecimento são utilizados de maneira mais adequada por pesquisadores quando se tem atrelado aos dados o uso de vocabulários controlados. Dessa forma, o uso de ontologias na construção ou reestruturação de repositórios, ou das bases de dados que disponibilizam dados primários, propicia uma melhor recuperação e utilização desses dados.

Neste contexto, usar as tecnologias da Web semântica, com melhorias semânticas na *e-Science,* traz potencialidades para melhor uso dos dados (BUCHAN; BISHOP, 2011; FOX; HENDLER, 2011).

Dessa forma, este artigo teve como objetivo a caracterização de três sítios que têm melhorias semânticas e uso de ontologias em suas páginas, verificando a correlação entre a melhoria semântica e a *e-Science*.

A metodologia utilizada consistiu no levantamento bibliográfico e revisão de literatura para discussão do tema. A partir da caracterização e da definição das quatro camadas da Web semântica (camada de base, camada sintática, camada de dados e camada ontológica) descritas, foram analisados três sítios com melhorias semânticas apresentados e localizadas no sítio da zoo.uk, da editora da Sociedade Real de Química, apresentada na literatura (GINSPARG, 2011) e da revista científica *Nature,* localizada a partir de verificação das melhorias semânticas em seu sítio.

Com o crescimento e o desenvolvimento de estudos e aplicações dessa tendência, torna-se fundamental a apresentação e orientação de métodos e modelos desenvolvidos para a compreensão e participação da comunidade científica.

2 WEB SEMÂNTICA

A Web semântica, termo apresentado por Tim Bernes-Lee em 2001, foi criada para representar um projeto do *World Wide Web Consortium* (W3C), que tinha o intuito de estruturar as páginas da Web de forma que as informações tivessem significados (BERNERS-LEE; HENDLER; LASSILA, 2001).

> O projeto da Web Semântica tem como ponto fundamental a criação de uma nova estrutura de armazenamento de dados. O ponto principal está na separação da apresentação do conteúdo e do conteúdo da estrutura, tratando as unidades atômicas de uma informação como componentes independentes. (SANTARÉM SEGUNDO, 2014, p. 3.865).

Pesquisadores da Web semântica, com influência da inteligência artificial e pelo aumento do volume de dados disponibilizados na Internet, têm pesquisas direcionadas aos aspectos formais das linguagens de representação semântica, como a XML, e no desenvolvimento de aplicações semânticas de uso geral. Essas linguagens estão sendo padronizadas, e as comunidades estão usando-as para a construção e uso de ontologias (FOX; HENDLER, 2011).

As ontologias e a interoperabilidade contribuem nesse contexto, propiciando significados e proporcionando a interação e o compartilhamento de dados e informações. A Web semântica é baseada em tecnologias interoperáveis e infraestrutura que possibilita aos computadores integrarem e processarem informações de acordo com seu significado (CHOWDHURY, G.; CHOWDHURY, S., 2007).

Fox e Hendler (2011) destacam que a ciência está cada vez mais dependente de dados. Tecnologias semânticas para a melhoria semântica estão sendo desenvolvidas em áreas como Ecologia e Física Solar-Terrestre, ciências com disponibilidade de grandes Datasets para a comunidade científica. As melhorias semânticas com uso de linguagens de representação e ontologias estão ganhando impulso, pois "A necessidade de mais semântica na e-Science também advém, em parte, dos desafios cada vez mais distribuídos e multidisciplinares da pesquisa moderna." (FOX; HENDLER, 2011, p. 161).

3 *e-SCIENCE* E A COMUNICAÇÃO CIENTÍFICA

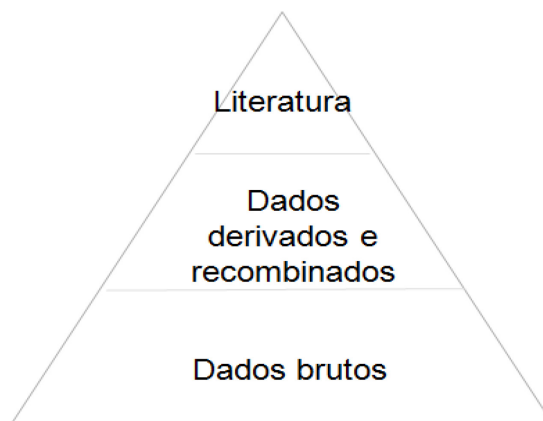
A comunicação científica é imprescindível para o desenvolvimento da ciência, produção de novos conhecimentos e reconhecimento dos estudos desenvolvidos por pesquisadores. As comunidades científicas produzem dados e informações que são registrados através das publicações científicas. "São conteúdos especializados resultantes de informações e procedimentos técnicos, talentos e restrições, experiência; acumulada, atitudes, normas e valores." (BERTO, 2003, p. 3).

A comunicação científica consiste em assegurar o intercâmbio de informações sobre os trabalhos em andamento e colocar os cientistas em contato entre si e com as pesquisas concomitantes (LE COADIC, 2004).

O registro e a comunicação científica têm o intuito de disseminar o conhecimento produzido e proporcionar visibilidade, reconhecimento e credibilidade ao pesquisador, à

pesquisa e à sua Instituição. São uma forma de colaboração em larga escala, com a possibilidade de reprodução de resultados das pesquisas e experimentos, além de oferecer evidências para a qualidade do trabalho científico (LYNCH, 2011). A comunicação está situada no coração da ciência (MEADOWS, 1999).

O registro científico disponibilizado na Internet pode conter o texto completo e os dados primários de pesquisa coletados pelo autor. A figura 1 mostra como todos os dados coletados podem ser integrados e concomitantemente "[...] unificar todos os dados científicos e toda a literatura para criar um mundo em que os dados e a literatura possam interagir." (TOLLE; TANSLEY; HEY, 2011, p. 24).

Figura 1 – Todos os dados científicos online.

Fonte: Adaptado de TOLLE; TANSLEY; HEY, 2011, p. 25.

Na figura 1 verifica-se que os dados brutos estão na base da pirâmide, destacando-se pela grande quantidade. Os dados podem ser combinados, recombinados e usados por áreas multidisciplinares, ser unificados, integrados e interoperados em rede. Dessa forma, na operacionalização do trabalho científico, o pesquisador tem a possibilidade de:

- ➔ Acessar os dados primários coletados por outros pesquisadores;
- ➔ Fazer a análise desses dados;
- ➔ Localizar e acessar os documentos publicados relacionados ao tema;
- ➔ Voltar aos dados primários;
- ➔ Desenvolver ou refazer sua análise, formular suas hipóteses e continuar sua pesquisa.

Essa integração facilita a produtividade científica por meio do aumento da velocidade com que a informação pode ser produzida e tratada (LYNCH, 2011), com a viabilidade de analisar e tratar dados já coletados, concentrando esforços na análise dos dados já existentes e compartilhados, reduzindo trabalho, custos e tempo.

Para Mueller (2000, p. 25), os resultados das pesquisas alcançados por determinados pesquisadores "[...] são frequentemente retomados por outros cientistas, teóricos ou aplicadores, que dão continuidade ao estudo, fazendo avançar a ciência ou produzindo tecnologias ou produtos neles baseados."

De acordo com Lynch (2011), o registro científico poderia disponibilizar os dados primários que foram utilizados para seu desenvolvimento, para que outros pesquisadores pudessem reaplicá-los para reproduzir novos resultados.

Jim Gray desejava aplicar a computação ao aumento da produtividade acadêmica e acelerar o ritmo das descobertas e inovações para os pesquisadores (DIRKS, 2011). A visão de Gray era que os recursos acadêmicos (textos, base de dados e outros recursos associados) fossem navegáveis e interoperáveis de forma integrada (GINSPARG, 2011).

A ideia do quarto paradigma da pesquisa científica emergiu com Gray e foi definida por Lynch (2011) de ciência intensiva em dados, que apresenta em suas discussões como o quarto paradigma é aplicado ao campo da comunicação acadêmica.

Gray (2007) apresenta uma evolução da ciência onde relaciona os quatro paradigmas da ciência:

Primeiro paradigma: mil anos atrás, a ciência era empírica, com a descrição dos fenômenos naturais;

Segundo paradigma: há poucos séculos, desenvolveu-se a ciência teórica, com o uso de modelos e generalizações, mostrando, como exemplo, as Leis de Newton, Kepler e as equações de Maxwell;

Terceiro paradigma: nas últimas décadas, apareceu o ramo computacional, com a simulação de fenômenos complexos, gerando uma grande quantidade de dados e mostrando o caminho para o quarto paradigma, destacado por ele (GRAY, 2007) como *e-Science;*

Quarto paradigma: a ciência do século XXI é apresentada como exploração de dados, a *e-Science,* que unifica os três paradigmas anteriores (teoria, experimento e simulação), destacando como características:

- Grandes quantidades de dados capturados por instrumentos ou gerados por simulações e processados por *softwares;*
- Informação e/ou conhecimento armazenado em computadores;
- Cientista analisa base de dados e arquivos por meio de gerenciamento de dados e estatísticas.

Segundo Medeiros e Caregnato (2012), a *e-Science* é uma infraestrutura que:

- Visa permitir aos cientistas e pesquisadores terem acesso a dados científicos primários distribuídos;
- Facilita o gerenciamento, o compartilhamento e a colaboração desses dados;
- Possibilita aos cientistas e pesquisadores de diversos ramos terem acesso a conteúdo já mapeado, contribuindo para o avanço da ciência.

A *e-Science* se desenvolveu devido à necessidade de enfrentar o "dilúvio" de dados em lugares diferentes. Na literatura é apresentada com nomes diferentes, como quarto paradigma, ciência orientada por dados, computação fortemente orientada a dados, ciberinfraestrutura e Dos dados ao conhecimento (CESAR JUNIOR, 2011).

Medeiros e Caregnato (2012, p. 315) destacam que:

> [...] *e-Science* altera consideravelmente a maneira com que os cientistas realizam seu trabalho, as ferramentas que utilizam, os tipos de problemas que abordam e a natureza da documentação e da publicação que resulta da sua pesquisa.

> Todas as capacidades atualmente necessárias à *e-Science* – integração de dados, fusão e mineração; desenvolvimento de fluxos de trabalho, orquestração e execução; captura da proveniência, linhagem e qualidade dos dados; validação, verificação e confiança na autenticidade dos dados; e adequação ao propósito – precisam de representação e mediação para que a *e-Science* possa se tornar realmente intensiva em dados. (FOX; HENDLER, 2011, p. 161).

A "ciência intensiva em dados", um dos componentes da *e-Science,* deve avançar para permitir o acesso aos dados pelos cientistas e pesquisadores que não fazem parte das equipes dos grandes projetos e "[...] permitir maior integração de fontes e prover interfaces para quem é especialista em ciência, mas não em computação e administração de dados." (FOX; HENDLER, 2011, p. 159).

Grandes projetos, como Pesquisa Celeste Digital *Sloan,* Grande *Colisor* de Hádrons (GCH) (TOLLE; TANSLEY; HEY, 2011), Projeto Australiano de radiotelescópios ASKAP e o conjunto de telescópios astronômicos Pan-STARRS, estão gerando *petabytes* de dados que são analisados por cientistas do mundo inteiro, em laboratórios diferentes e que falam línguas diferentes.

Os dados coletados pelos pesquisadores nas diversas áreas do conhecimento estão sendo considerados como informação científica e precisam ser tratados de forma a viabilizar a sua organização, recuperação e difusão para auxiliar a pesquisa colaborativa.

Santos e Sant'Ana (2002) destacam que dado é conceituado como "[...] um elemento básico, formado por signo ou conjunto finito de signos que não contém, intrinsecamente, um componente semântico, mas somente elementos sintáticos."

Viabilizar melhorias semânticas para que os dados primários sejam processados automaticamente por máquinas e melhor recuperados por humanos se torna essencial. A partir desse ponto, serão abordadas algumas possibilidades que a *e-Science* semântica oferece para responder às discussões levantadas.

4 *e-SCIENCE* SEMÂNTICA

Fox e Hendler (2011) destacam o desenvolvimento de metodologias e ferramentas baseadas em semântica na ambiência da *e-Science* e no desenvolvimento de sua infraestrutura.

Segundo Fox e Hendler (2011), a necessidade de mais semântica na *e-Science* advém de fatores como os desafios mais distribuídos e multidisciplinares da pesquisa moderna. Os autores destacam que

> À medida que crescem o volume, a complexidade e a heterogeneidade dos recursos de dados, os cientistas precisam cada vez mais de novas capacidades, que dependem de novas abordagens "semânticas" (por exemplo, sob a forma de **ontologias)** [...] (FOX; HENDLER, 2011, p. 159, grifo dos autores).

Para os pesquisadores que utilizarão dados provenientes de domínios de conhecimentos diferentes, torna-se mais tangível utilizá-los e reutilizá-los quando embasados no significado dos dados. Esse conhecimento pode ser proporcionado por meio de melhorias semânticas e do uso de ontologias que definem especificações de conceitos e suas inter-relações.

De acordo com Santarém Segundo (2014, p. 3.866),

> Utilizar ontologias e suas relações é uma das maneiras de se construir uma relação entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação.

Segundo o consórcio W3C (2004), ontologia é a definição dos termos utilizados para descrição e representação de uma área do conhecimento. As ontologias definem o vocabulário em uma área específica de domínio e fornecem infraestrutura para integrar bases de conhecimento.

A Web semântica está baseada em tecnologias interoperáveis e infraestrutura que permite computadores integrarem e processarem informações de acordo com seu significado e uso (CHOWDHURY, G.; CHOWDHURY, S., 2007).

A práxis das tecnologias da Web semântica aparecem em áreas da *e-Science* como a Física Solar-Terrestre, Ciências Marítimas e Oceânicas, como o Projeto de Interoperabilidade de Metadados Marinhos (MMI), Ecologia, serviços de saúde e ciências da vida entre outros (FOX; HENDLER, 2011). Os envolvidos nos esforços "[...] defendem uma interoperabilidade que se afaste do nível do elemento do dado, o nível sintático, e avance para um nível científico mais alto, o nível semântico." (FOX; HENDLER, 2011, p. 162).

Essa necessidade da semântica advém das características da pesquisa moderna que está, atualmente, mais distribuída e multidisciplinar, com contribuição de pesquisadores de diversos laboratórios e centros de pesquisa (FOX; HENDLER, 2011). Essas características contribuem e/ou influenciam outros domínios no percurso das pesquisas, no uso, reúso, análise e na integração dos dados (FOX; HENDLER, 2011).

> Os desenvolvedores de infraestruturas para *e-Science* precisam cada vez mais de metodologias, ferramentas e *middleware* baseados em semântica. Com isso, podem facilitar a modelagem de conhecimento científico, a verificação de hipóteses baseadas em lógica, a integração dos dados semânticos, a composição de aplicações e a integração de descoberta de conhecimento e análise de dados para os diferentes domínios e sistemas mencionados acima, para uso por cientistas, estudantes e, cada vez mais, não especialistas (FOX; HENDLER, 2011, p. 160).

Os atuais desafios da *e-Science* semântica são equilibrar a expressividade da representação semântica com a complexidade da definição de termos utilizados por especialistas e implementar os sistemas resultantes (FOX; HENDLER, 2011). Dessa forma, o

uso de ontologias contribui para conhecer o significado dos dados, pois "[...] o dado em si, se transmitido ou registrado fora do seu contexto, pouco ou nada pode representar em termos de significado." (SANTOS; SANT'ANA, 2002).

A interoperabilidade é outro aspecto importante. Permite a troca de dados e informações entre sistemas, bases de dados, repositórios e, assim, facilita a difusão eficiente entre os conteúdos.

Embora taxonomia e ontologias tenham tipos de relacionamentos diferentes entre os conceitos, a taxonomia com relação hierárquica e a ontologia com relação em rede, ambas representam a estrutura conceitual de domínios. Segundo definição da W3C (2004), as ontologias, sob o aspecto estrutural, são definidas como um conjunto de definições legíveis por máquina, que criam uma taxonomia de classes e subclasses e os relacionamentos entre elas.

Dessa forma, apresentaremos três sistemas que contêm melhorias semânticas e que utilizam ontologias e taxonomias em seus sítios.

5 COLETA E ANÁLISE DOS RESULTADOS

Conforme apresentado neste artigo, as iniciativas de melhorias semânticas estão sendo difundidas entre as áreas do conhecimento para melhor representação e recuperação de recursos na Web.

Observamos a presença de elementos semânticos nas camadas da estrutura da Web semântica, conforme se seguem nos três exemplos apresentados. As camadas observadas foram: camada de base, sintática, de dados e ontológica.

Os elementos observados nas estruturas das camadas dos sítios dos exemplos foram:

- URI *(Uniforme Resource Locator):* define a localização do recurso;
- XML *(eXtensible Markup Language):* linguagem de descrição para estruturação de documentos;
- XML *schema:* define elementos e atributos que podem aparecer no documento;
- RDF *(Resource Description Framework):* uma aplicação da linguagem XML e tem a capacidade de associar a tríade recurso, propriedade e valor (SANTARÉM SEGUNDO, 2015);

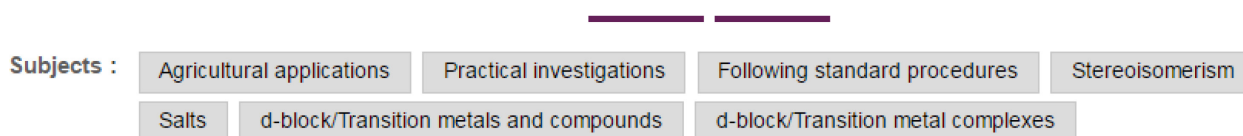
ARTIGO

- RDF *schema:* extensão do RDF fornece um vocabulário de dados para o RDF (W3C, 2014);
- Ontologia: é a definição dos termos utilizados para descrição e representação de uma área do conhecimento (W3C, 2004);
- OWL *(Web Ontology Language):* linguagem de representação que pode ser explorada por programas de computador (W3C, 2004).

A publicação *Molecular BioSystems,* da Sociedade Real de Química, identificada como editora, tem o aperfeiçoamento semântico em seu HTML, pois destaca termos do texto que estão listados em bases de dados de terminologia de química e os conecta aos verbetes das bases de dados externas, conecta termos de ontologias de genes, sequências e células (GINSPARG, 2011). O sistema aplica em sua estrutura a abordagem semântica da *e-Science.*

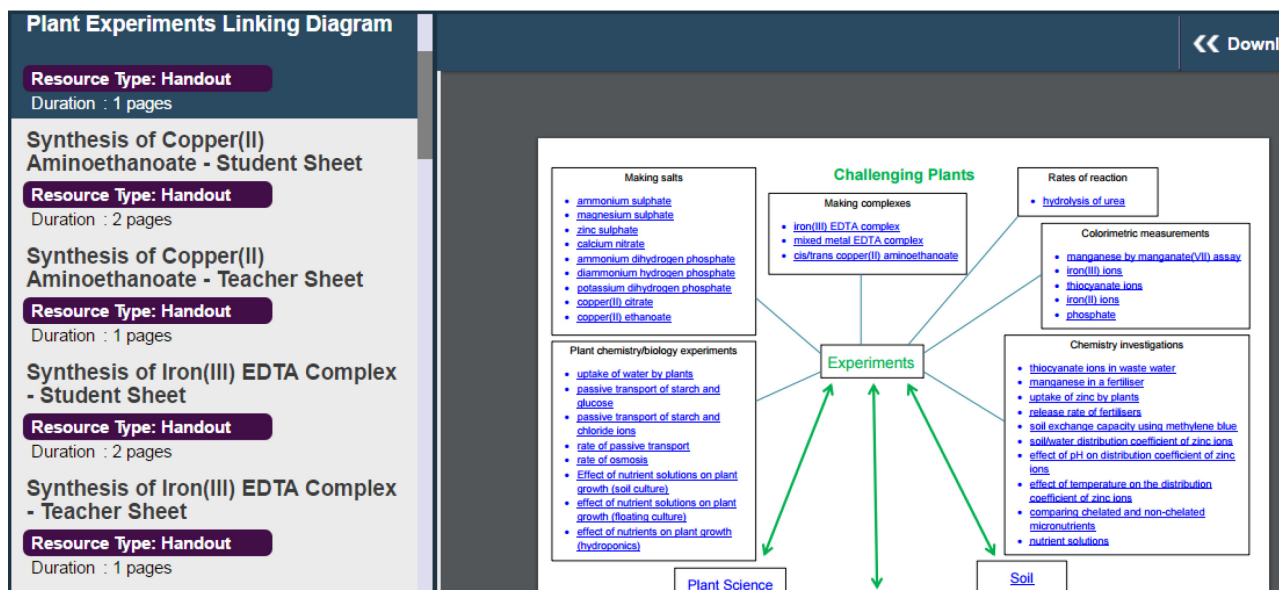
Usamos, como exemplo, um artigo da base para verificação e caracterização, conforme a figura 2:

Figura 2 – Descritores do artigo com melhoria semântica.

Subjects : Agricultural applications | Practical investigations | Following standard procedures | Stereoisomerism | Salts | d-block/Transition metals and compounds | d-block/Transition metal complexes

Fonte: *LEARN CHEMISTRY*, c2016.

Na estrutura do documento apresentado da *Learn Chemistry,* conforme pode ser observada na figura 2, os descritores estão apresentados logo abaixo da descrição do conteúdo do artigo e dão o link para outros documentos relacionados com os descritores destacados. Os descritores, quando acionados, remetem para outra página com os documentos relacionados ao seu assunto dentro da base de dados, deixando a busca e a recuperação mais dinâmica e facilitada para o usuário.

Observamos, na figura 3, que, além dos destaques dos descritores, o sítio disponibiliza, junto ao artigo, os componentes apresentados, suas estruturas, fórmulas e aplicações, interligando os dados e textos e deixando a recuperação mais dinâmica e eficiente, sem que o usuário precise buscar os dados e as informações apresentadas em outras fontes ou fazer uma nova busca na própria base de dados:

Figura 3 – Dados apresentados relacionados às estruturas do artigo.

Fonte: *LEARN CHEMISTRY*, c2016.

Verificamos a presença de elementos nas camadas da estrutura da Web semântica, como seguem nos exemplos apresentados.

- Na camada de base: utiliza URI, proporcionando a localização dos recursos;
- Na camada sintática: utiliza a linguagem XML, que possibilita a descrição das estruturas dos dados e das informações. Utiliza também os *names spaces*;
- Na camada de dados: utiliza o RDF;
- Na camada ontológica, que é responsável pela semântica dos dados, utiliza ontologia de bases de dados de química.

Um exemplo de enriquecimento semântico em artigos é apresentado por David Shotton, da Universidade de Oxford, conforme apresentado na figura 4:

Figura 4 – Artigo com melhorias semânticas.

turn all highlighting on | date | disease | habitat | institution | organism | person | place | protein | taxon

Top | Abstract | Author Summary | Introduction | Methods | Results | Discussion | Supporting Information | Acknowledgements | References | Data Fusion Supplements

SEMANTICALLY ENHANCED VERSION OF A RESEARCH ARTICLE FROM PLOS NEGLECTED TROPICAL DISEASES

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slums

document summary

Renato B. Reis [1#], Guilherme S. Ribeiro [1#], Ridalva D. M. Felzemburgh [1], Francisco S. Santana [1,2], Sharif Mohr [1], Astrid X. T. O. Melendez [1], Adriano Queiroz [1], Andréia C. Santos [1], Romy R. Ravines [3], Wagner S. Tassinari [3,4], Marília S. Carvalho [3], Mitermayer G. Reis [1], Albert I. Ko [1,5*]
1 Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Ministério da Saúde, Salvador, Brazil 2 Secretária Estadual de Saúde da Bahia, Salvador, Brazil 3 Escola Nacional da Saúde Pública, Fundação Oswaldo Cruz, Ministério da Saúde, Rio de Janeiro, Brazil 4 Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil 5 Division of International Medicine and Infectious Diseases, Weill Medical College of Cornell University, New York, New York, United States of America

Abstract

Background

Leptospirosis has become an urban health problem as slum settlements have expanded worldwide. Efforts to identify interventions for urban leptospirosis have been hampered by the lack of population-based information on *Leptospira* transmission determinants. The aim of the study was to estimate the prevalence of *Leptospira* infection and identify risk factors for infection in the urban slum setting.

Methods and Findings

We performed a community-based survey of 3,171 slum residents from Salvador, Brazil. *Leptospira* agglutinating antibodies were measured as a marker for prior infection. Poisson regression models evaluated the association between the presence of *Leptospira* antibodies and environmental attributes obtained from Geographical Information System surveys and indicators of socioeconomic status and exposures for individuals. Overall prevalence of *Leptospira* antibodies was 15.4% (95% confidence interval [CI], 14.0–16.8). Households of subjects with *Leptospira* antibodies clustered in squatter areas at the bottom of valleys. The risk of acquiring *Leptospira* antibodies was associated with household environmental factors such as

Fonte: Disponível em: <http://svn.code.sf.net/p/enhancedplospaper/code/trunk/paper/index.html#pntd-0000228-t001>. Acesso em: 11 jun. 2016.

Na figura 4, podemos verificar os destaques em cores relacionados às melhorias semânticas no artigo. No alto da tela, são apresentadas classes de termos que estão interligados e identificados com uma cor diferente. Ao serem acionados, marcam as palavras que são correspondentes aos termos no texto deixando as palavras com a mesma cor deste termo. Por exemplo, a palavra *Urban Sluns* com destaque em verde no título e no corpo do texto.

Outro aspecto de melhoria semântica é apresentado na figura 5. Verificamos que, ao clicar sobre as palavras do texto relacionado aos termos *"organism",* com destaque em azul, abre-se um vocabulário com as definições e os relacionamentos do termo:

ARTIGO

Figura 5 – Vocabulário controlado dos termos em destaque no texto.

We performed a community-based sur... ptospira agg...
regression models evaluated the assoc... and environ...
indicators of socioeconomic status and... ospira antib...
with *Leptospira* antibodies clustered in... cquiring *Lep*...
residence in flood-risk regions with op... 1.75) and pr...
the presence of chickens (1.26, 1.05–1... 1.03–1.50) ...
income was associated with an 11% (9...

chicken
NamebankID: 1903523
LSID: urn:lsid:ubio.org:namebank:1903523 (view metadata)

Classification: (According to: *Birds: Zoonomen*)

Aves
Neornithes
Neoaves
Passerae
Passerimorphae
Ciconiiformes
Charadrii
Charadriides
Scolopacida
Scolopacoidea
Scolopacidae
Arenaria Brisson 1760
Arenaria interpres (Linnaeus) 1758

Alternate Classifications:
Birds: AOU 83
Birds: AOU 98
Birds: James Lee Peters
Birds: Morony, Bock, and Farrand
Birds: Nomina Avium Mundi
Birds: Sibley and Monroe
Birds:Howard and Moore (edition 2)
Chapin, Birds of the Belgian Congo
Howard and Moore, 3rd Edition

Fonte: Disponível em: <http://www.ubio.org/browser/details.php?namebankID=1903523>. Acesso em: 11 jun. 2016.

Na análise do sítio, verificamos que também são usadas as camadas da Web semântica. Na estrutura da Web semântica, verificamos:

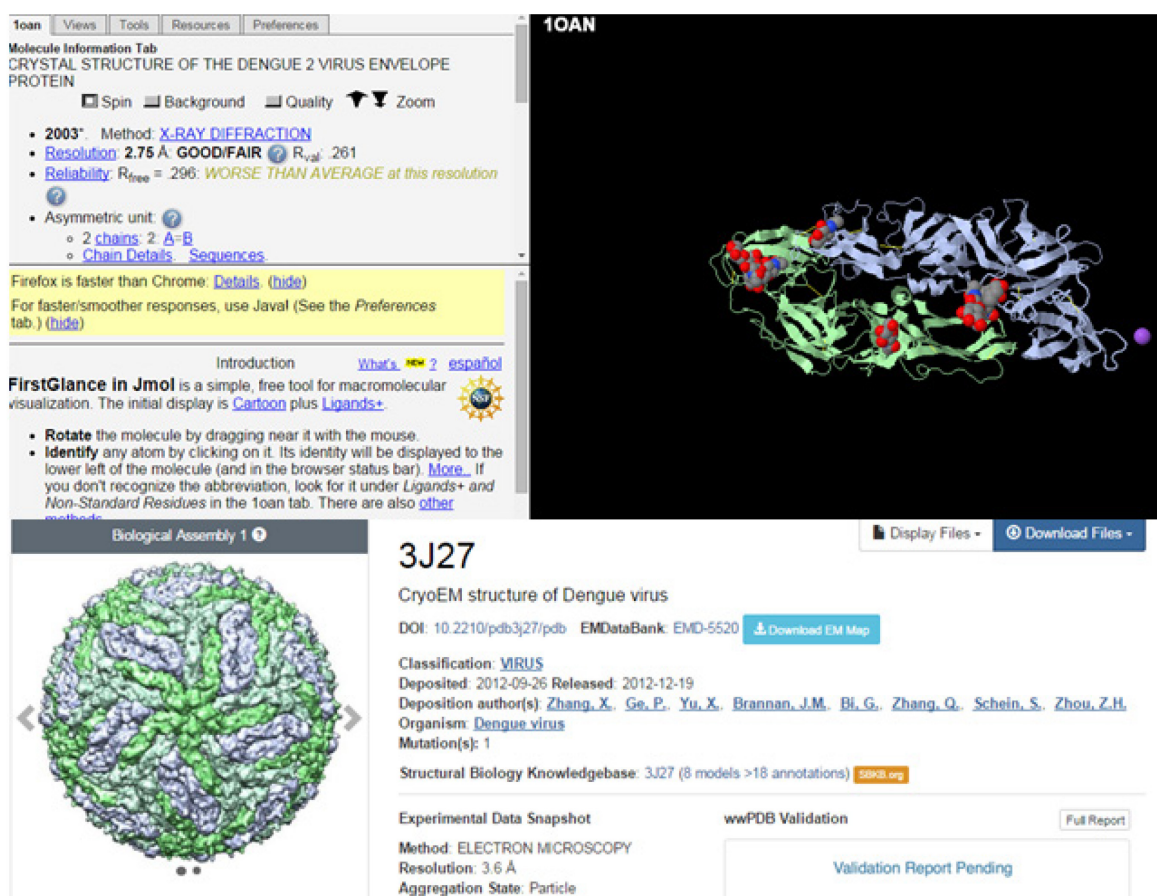
- Na camada de base: utiliza URI, proporcionando a localização dos recursos;
- Na camada sintática: utiliza a linguagem XML, que possibilita a descrição das estruturas dos dados e das informações. Utiliza também os *names spaces;*
- Na camada de dados: utiliza RDF;
- Na camada ontológica, que é responsável pela semântica dos dados, está representada pela OWL – ontologia do Protégé.

A revista *Nature* tem aperfeiçoamento semântico em sua estrutura. Na observação de um de seus artigos, verificamos que os dados estão integrados:

Figura 6 – Aperfeiçoamento semântico da revista *Nature*.

Fonte: Disponível em: <http://www.nature.com/ni/journal/v16/n2/full/ni.3058.html#accessions>. Acesso em: 11 jun. 2016.

Conforme observado nas figuras 6 e 7, a seguir, o artigo apresenta a discussão teórica dos autores, suas tabelas, gráficos, imagens e métricas, dando acesso às ilustrações em tamanho original, com visualização em 3D, e aos dados coletados pelos autores, em Excel, para montagem das tabelas e gráficos. O usuário navega pelo sítio, tem a possibilidade de verificar as informações sobre os autores, outros artigos de sua autoria, documentos relacionados aos assuntos, dados originais das pesquisas e acesso a algumas das referências citadas. Dessa forma, o uso da Web semântica no portal integra os dados e as informações, facilitando a descoberta de novos conteúdos e maior eficiência e relevância na recuperação:

Figura 7 – Estrutura do vírus da dengue.

Fonte: Disponível em: <http://bioinformatics.org/firstglance/fgij//fg.htm?mol=1OAN>. Acesso em: 11 jun. 2016.

No processo de enriquecimento semântico do portal, inferimos que há aplicação de algoritmos para descobrir relações e interligações entre os conceitos. Na análise do sítio, verificamos também que se utilizam as camadas da Web semântica no portal. Na estrutura da Web semântica, verificamos:

- Na camada de base: utiliza-se URI, proporcionando a localização dos recursos;
- Na camada sintática: utiliza-se a linguagem XML, que possibilita a descrição das estruturas dos dados e das informações. Utiliza também os *name spaces;*
- Na camada de dados: utiliza RDF para prover a interoperabilidade;
- Na camada ontológica, que é responsável pela semântica dos dados, está representada pela taxonomia da *National Center Biotechnology Information* (NCBI).

Quadro 1 – Identificação das camadas nos três sítios.

Camadas	Sítio 1	Sítio 2	Sítio 3
Base	URI	URI	URI
Sintática	XML	XML	XML
Dados	RDF	RDF	RDF
Ontológica	Base de dados em química	OWL Protegè	Taxonomia NCBI

Fonte: Elaborado pelos autores.

Podemos observar, no quadro 1, os três exemplos dos sítios com a implementação do aperfeiçoamento semântico. O sítio da *Learn Chemistry,* identificado como sítio 1, o sítio da zoo.uk, identificado como sítio 2, e o sítio da *Nature,* identificado como sítio 3. Nestes três exemplos a implementação do aperfeiçoamento semântico em seus sítios, na tentativa de trazer mais significado e integração em seus dados e em suas informações, está presente, apresentando as características da Web semântica na *e-Science,* aperfeiçoando, assim, a recuperação dos dados utilizados na pesquisa.

6 CONSIDERAÇÕES FINAIS

A ciência, em suas ambiências, está sendo afetada pelo "dilúvio" de dados, e, com isso, emerge o Quarto paradigma da ciência: ciência orientada a dados. A grande quantidade de dados disponíveis está crescendo rapidamente e excede a capacidade da extração e da análise desses dados sem a ajuda de aparatos tecnológicos. Projetos com melhorias semânticas surgem para melhorar recuperação, compartilhamento, entendimento e reúso dos dados.

A *e-Science* traz oportunidades para o avanço da ciência, embasada pela colaboração multidisciplinar e internacional, melhorias semânticas, a integração e o compartilhamento de dados.

A estrutura e o fluxo da publicação e comunicação científica, permeados pelo desenvolvimento das TIC, da computação e de redes eletrônicas, bem como pelos portais com melhorias semânticas e interoperabilidade, estão propícios a novas práxis que os ampliam, diversificam e os tornam mais rápidos e abrangentes, viabilizando novos parâmetros no âmbito acadêmico/científico.

Verificamos que os envolvidos nos trabalhos de melhorias semânticas apresentados demonstram esforços para ter interoperabilidade em nível semântico, além da integração de dados para o uso multidisciplinar com a utilização de diferentes instrumentos, mas com algo em comum: o uso da proposta de melhoria semântica em suas estruturas.

Embora haja algumas iniciativas para o desenvolvimento da *e-Science* se tornar semântica, ainda há muito para se pesquisar e incrementar no que diz respeito ao desenvolvimento e uso de ferramentas tecnológicas e interação entre pesquisadores.

As possibilidades são muitas, e os desafios também. As iniciativas podem atingir seu potencial pleno à medida que as ferramentas e os pesquisadores estiverem engajados nesse paradigma da ciência, trabalhando para estabelecer um sistema em que dados e descobertas científicas possam ser disponibilizados, compartilhados e reutilizados de maneira mais eficiente.

REFERÊNCIAS

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web: a new formo f Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American,** May 2001. Disponível em: <http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf>. Acesso em: 10 jun. 2016.

BERTO, R. M. V. S. **Novas práticas de comunicação e produção de publicações científicas.** Disponível em: <http://reposcom.portcom.intercom.org.br/bitstream/1904/5281/1/EDOCOM_BERTO.pdf>. Acesso em: 12 abr. 2006.

BUCHAN, I.; BISHOP, J. W. C. Uma abordagem unificada da modelagem de serviços de saúde com uso intensivo em dados. In: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 113-119.

CESAR JUNIOR, R. M. Apresentação à edição brasileira. In: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 7-8.

CHOWDHURY, G. G.; CHOWDHURY, S. The semantic web. *In*: ______. **Organizing information from the self to the web.** London: Facet Publishing, 2007. p. 111-129.

DIRKS, L. Introdução. *In*: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 185-186.

FOX, P.; HENDLER, J. e-Science semântica: o significado codificado na próxima geração de ciência digitalmente apropriada. *In*: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da e-Science. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 159-163.

GIL, A. C. **Métodos e técnicas de pesquisa social.** 2. ed. São Paulo: Atlas, 1989.

GINSPARG, P. O texto em um mundo centrado em dados. *In*: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 195-199.

GRAY, J. **e-Science:** a transformed scientific method. 2007. Disponível em: <http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt>. Acesso em: 18 ago. 2015.

______. Jim Gray on escience: a transformed scientific method. *In*: HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). **The fourth paradigm:** data-intensive scientific discovery. Washington: Microsoft Research, 2009. Disponível em: <http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf>. Acesso em: 05 maio 2015.

JACOB, E. K. Ontologies and the semantic web. **Bulletin for the American Society for Information Science and Technology,** v. 29, n. 4, p. 19-22, abr./maio 2003. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/bult.283/epdf>. Acesso em: 6 Jul. 2015.

LE COADIC, Y. F. **A ciência da informação.** 2. ed. Brasília, DF: Briquet de Lemos/Livros, 2004.

LYNCH, C. O quarto paradigma de Jim Gray e a construção do registro científico. *In*: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 187-193.

MEADOWS, A. J. **A comunicação científica.** Brasília, DF: Brinquet de Lemos/Livros, 1999.

MEDEIROS, J. C.; CARAGNATO, S. E. Compartilhamento de dados e e-Science: explorando um novo conceito para a comunicação científica. **Liinc em Revista,** Rio de Janeiro, v. 8, n. 2, p. 311-322, set. 2012. Disponível em: <http://www.ibict.br/liinc>. Acesso em: 8 abr. 2015.

MORAES, S. H. M. H.; BELLUZZO, R. C. B. Informação, conhecimento & gestão de projetos: da sistematização de princípios à aplicação em ambientes acadêmicos para captação de recursos à pesquisa. *In*: VIDOTTI, S. A. B. G. (Org.). **Tecnologia e conteúdos informacionais:** abordagens teóricas e práticas. São Paulo: Polis, 2004. p. 77-94.

MUELLER, S. P. M. A ciência, o sistema de comunicação científica e a literatura científica. *In*: CAMPELLO, B. C.; CENDÓN, B. V.; KREMER, J. M. (Org.). **Fontes de informação para pesquisadores e profissionais.** Belo Horizonte: UFMG, 2000. p. 21-34.

RIBES, D.; LEE, C. P. Sociotechnical studies of cyberinfrastructure and e-research: current themes and future trajectories. **Computer Supported Cooperative Work,** v. 19, n. 3-4, p.

231-244, 2010. Disponível em: <http://www.davidribes.com/storage/Ribes%20Lee%20-%20Cyberinfrastructure%20Studies.pdf>. Acesso em: 6 maio 2012.

SANTARÉM SEGUNDO, J. E. Web semântica: introdução a recuperação de dados usando *sparql. In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais eletrônicos...** Belo Horizonte: UFMG, 2014. p. 3863-3882. Disponível em: <http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8>. Acesso em: 1º maio 2015.

______. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais eletrônicos...** João Pessoa: UFPB, 2015. Disponível em: <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/3149/1193 >. Acesso em: 6 jun. 2016.

SANTOS, P. L. V. A. C.; SANT’ANA, R. C. G. Transferência da informação: análise para valoração de unidades de conhecimento. **DataGramaZero:** Revista de Ciência da Informação, v. 3, n. 2, abr. 2002. Disponível em: <http://www.dgz.org.br/abr02/Art_02.htm>. Acesso em: 1 maio 2015.

TOLLE, K.; TANSLEY, S.; HEY, T. Jim Gray e a eScience: um método científico transformado. *In*: HEY, T.; STEWARD, T.; TOLLE, K. (Org.). **O quarto paradigma:** descobertas científicas na era da eScience. Tradução Leda Beck. São Paulo: Oficina de textos, 2011. p. 17-29.

W3C. **OWL Web Ontology Language.** [Cambridge], 2004. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 10 ago. 2015.

______. **RDF Schema 1.1**. [Cambridge], 2014. Disponível em: <http://www.w3.org/TR/rdf-schema/>. Acesso em: 8 fev. 2016.

ARTIGO

SOBRE OS AUTORES

Elizabete Cristina de Souza de Aguiar Monteiro
Mestranda em Ciência da Informação pela Universidade Estadual Paulista (Unesp-Marília).
E-mail: beteaguia@yahoo.com.br

Ricardo Cesar Gonçalves Sant'Ana
Professor do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (Unesp-Marília).
E-mail: ricardosantana@marilia.unesp.br

José Eduardo Santarém Segundo
Docente e coordenador do Curso de Graduação em Ciências da Informação e da Documentação e Biblioteconomia, da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, da Universidade de São Paulo (USP); Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/Marília na linha de Informação e Tecnologia.
E-mail: santarem@usp.br

Recebido em: 26/04/2016; **Revisado em:** 31/05/2016; **Aceito em:** 16/06/2016.

Como citar este artigo

MONTEIRO, Elizabete Cristina de Souza de Aguiar; SANT'ANA, Ricardo Cesar Gonçalves; SANTARÉM SEGUNDO, José Eduardo. E-science semântica: integração dos dados na comunicação científica. **Informação em Pauta,** Fortaleza, v. 1, n. 1, p. 9-29, jan./jun. 2016.